Characterizing missed identifications and errors in latent fingerprint comparisons using eye-tracking data

Supporting Information

Contents

Glossary

This section defines terms and acronyms as they are used in this paper.

ACE-V	The prevailing method for latent print examination: Analysis, Comparison, Evaluation, Verification.
Analysis phase	The first phase of the ACE-V method. In these studies, the examiner assessed the latent and made a value determination before seeing the exemplar print.
Comparison phase (Comparison/Evaluation phase)	The second and third phases of the ACE-V method. In this test, there was no procedural demarcation between the Comparison and Evaluation phases of the ACE-V method; hence, this refers to the single combined phase during which both images were presented side-by-side.
Comparison determination	The determination of identification, exclusion, or inconclusive reached in the Comparison/Evaluation phase of ACE-V.
Conflict resolution	The process conducted when there is a difference of determinations or conclusions between examiners, generally when the initial examiner and verifier disagree.
Determination	The result of an examiner's decision: the Analysis phase results in a Value determination, and the Comparison/Evaluation phase results in a Comparison determination.
Exclusion	The comparison determination that the latent and exemplar fingerprints did not come from the same finger.
Exemplar	A fingerprint from a known source, intentionally recorded.
False negative (FN)	An erroneous exclusion of a mated image pair by an examiner.
False positive (FP)	An erroneous identification of a nonmated image pair by an examiner.
ID (identification)	The comparison determination that the latent and exemplar fingerprints originated from the same source.
Inconclusive	The comparison determination that neither identification nor exclusion is possible.
Insufficient	When referring to examiner determinations (response data), "Insufficient" responses include both latent NV determinations (Analysis phase) and inconclusive determinations (Comparison/Evaluation phase).
Latent (or latent print)	An image of a friction ridge impression from an unknown source. In North America, "print" is used to refer generically to known or unknown impressions. Outside of North America, an impression from an unknown source (latent) is often described as a "mark" or "trace," and "print" is used to refer only to known impressions (exemplars).
Mated	A pair of images (latent and exemplar) known a priori to derive from impressions of the same source (finger). Compare with "ID," which is an examiner's determination that the prints are from the same source.
Missed ID	Failure by an examiner to identify a mated pair that was identified by a consensus of other examiners.
Nonmated	A pair of images (latent and exemplar) known a priori to derive from impressions of different sources (different fingers and/or different subjects).
NV (No value)	The impression is not of value for identification and contains no usable friction ridge information.
Source	An area of friction ridge skin from which an impression is left. Two impressions are said to be from the "same source" when they have in common a region of overlapping friction ridge skin.
Sufficient	An examiner's assessment that the quality and quantity of information in a print (or image pair) justifies a specific determination (especially used with respect to identification).
True negative (TN)	An exclusion of a nonmated image pair by an examiner.
True positive (TP)	An identification of a mated image pair by an examiner.
Value determination	An examiner's determination of the suitability of an impression for comparison: Of Value (sometimes divided into value for identification (VID) vs. value for exclusion only (VEO)), or No Value (NV).
Verification	The final phase of ACE-V: the independent application of the ACE process by a subsequent examiner to either support or refute the conclusions of the original examiner.
VID	Determination based on the analysis of a latent that the impression is of value and is appropriate for potential identification if an appropriate exemplar is available. See also VEO and NV.
VEO	Value determination based on the analysis of a latent that the impression is of value for exclusion only and contains some friction ridge information that may be appropriate for exclusion if an appropriate exemplar is available. See also NV and VID.

Appendix SI-1 Fingerprint data description

	Mating	*Selection Group*	*Comparison sets*	*Mates*	*Nonmates*
Latent-exemplar	Mates	Low reproducibility mates (including erroneous exclusions)	22	22	-
		Unanimous identification	2	2	-
		Unanimous No Value or inconclusive	1	1	-
	Nonmates	Erroneous identification (actual)	6	-	6
		Erroneous identification (potential)	5	-	5
		Low reproducibility nonmates	5	-	5
		Unanimous exclusion	2	-	2
		Unanimous no value or inconclusive	2	-	2
	Total		45	25	20
Exemplar-exemplar	Mates	Obvious ID	8	8	-
	Nonmates	Similar pattern class	4	-	4
		Unrelated pattern class	6	-	6
	Total		18	8	10

Table S1. Basis for selection for the image pairs used in the study. Selection was based on responses in the earlier black box and/or white box studies [1, 2](4-40 previous responses per image pair, mean 12.9)

The experimental design balanced among competing objectives: to include a variety of image pairs, to collect conclusions from multiple examiners for every image pair, and to give each examiner image pairs with a similar distribution of difficulty and other attributes. To this end, of the 45 latent-exemplar image pairs, 15 were each assigned to 1/3 of the participants, and the remaining 30 were each assigned to 1/6 of the participants. These were assigned in blocks to control the assignment of attributes of image pairs. Fig S1 illustrates the design structure for all participants.

The study was designed assuming that examiners would generally complete 8-12 standard comparisons, and that many participants would not complete all assignments (in the end, the majority of participants completed all 15 assigned latent-exemplar comparisons: median 15, mean 12.1). The directed tasks and easy (exemplar-exemplar) comparisons were added based on the assumptions that they would be fast enough (< 1 minute each) that they would not perturb the counts of latent-exemplar comparisons, which were the main focus of the study.

The first four participants were each assigned only latent-exemplar comparisons, and none completed all assignments; feedback indicated that the unbroken series of difficult comparisons was overwhelming. All later participants received assignments in which the comparisons were grouped in this way:

- "A group" included 15 latent-exemplar image pairs and 6 exemplar-exemplar image pairs, each assigned to a third of all examiners. In the design, we planned for 40 examiners per image pair if 120 examiners participated (in the end, we had 121 examiners, and a mean of 39.6 A comparisons per participant). The three A groups (A1-A3) each contained 5 latent-exemplar comparisons and 2 exemplar-exemplar comparisons.

Each examiner was assigned to one A group, so all (e.g.) A1 comparisons were done by the same ~40 examiners, making it easier to do examiner-examiner comparisons (this was based on a lesson learned from the latent print black box study, in which random assignments made it challenging to assess examiner vs. image effects). The images were assigned to make the A groups as similar as possible in terms of difficulty: for example, each of the A groups contained one nonmated image pair that had received multiple exclusion responses, multiple No Value/inconclusive responses, and no ID responses in the White Box study.

- "B group" included 30 latent-exemplar image pairs and 18 exemplar-exemplar image pairs, each assigned to a sixth of all examiners. The six B groups (B1-B6) each contained 5 latent-exemplar comparisons and 2 exemplar-exemplar comparisons. The images were assigned to make the B groups as similar as possible in terms of difficulty.
- Directed tasks were each assigned to a third of all examiners (groups D1-D3). Each D group included 6-8 find-the-target tasks, up to 3 ridge-counting tasks, and up to 3 ridge-tracing tasks. The earlier assignments included 8 find-the-target tasks, 3 ridge-counting tasks, and 3 ridge-tracing tasks for each participant, but after the first 35 participants we decided that was excessive and dropped the directed tasks to 6 find-the-target tasks, 1 ridge-counting task, and 1 ridge-tracing task.

Each participant was randomly assigned one A group and two B groups. For each participant, the image pair in the A groups were randomized, then presented before the B groups (also randomized), so if they quit early, they still would complete the A comparisons even if some would not complete all B assignments. Directed tasks were interspersed (assigned as every even-numbered trial until all were assigned).

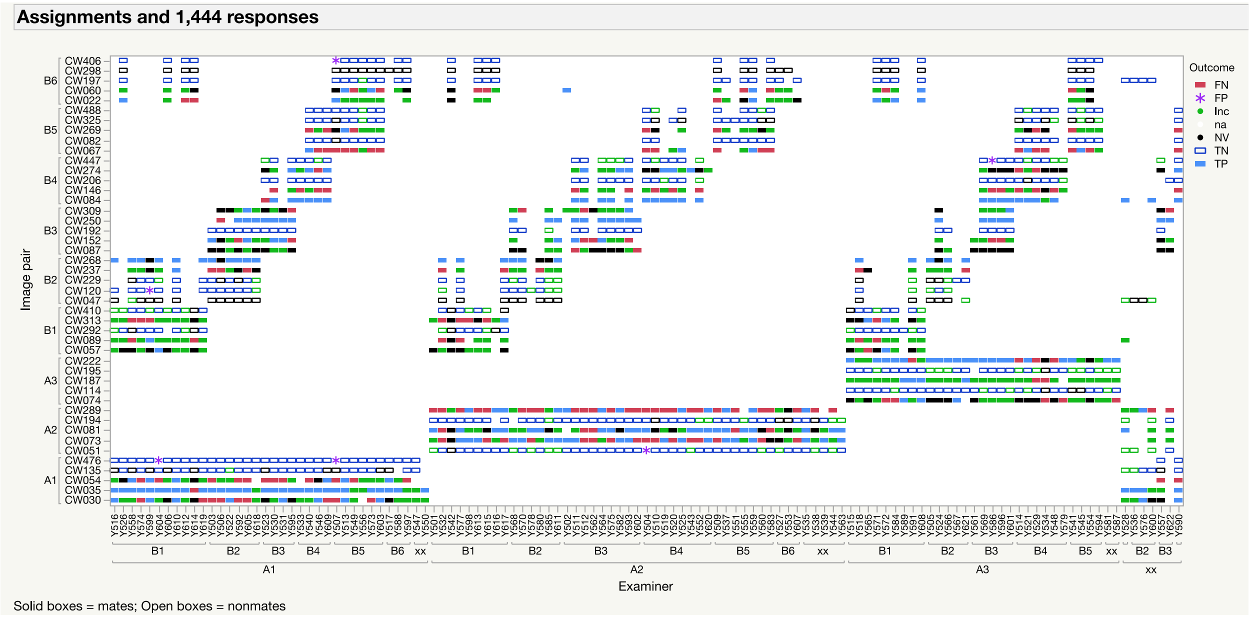


Fig S1. Design structure for the assignment of images to examiners, along with the outcome of each comparison. The images were grouped into blocks, and participants were randomly assigned to the A and B blocks. Each A image pair was assigned to 1/3 of participants, and each B image pair was assigned to 1/6 or participants. FN: False Negative (erroneous exclusion), FP: False Positive (erroneous identification), Inc (inconclusive), NV: No Value, TN: True Negative (correct exclusion), TP: True Positive (correct identification).

All images were scanned at 39.4 pixels per millimeter (1000 pixels per inch) in accordance with the prevailing standard (ANSI/NIST-ITL, 2013). Note that any references to pixel counts are integers @1000ppi, but metric equivalents are rounded.

Appendix SI-2 Distribution of Outcomes for All Image Pairs

Mating	*ImagePair*	*N Rows*	*N(TP)*	*N(TN)*	*N(FP)*	*N(FN)*	*N(Inc)*	*N(NV)*
Mates	CW022	29	6	-	-	3	16	4
Mates	CW030	41	10	-	-	9	15	7
Mates	CW035	42	34	-	-	1	7	-
Mates	CW054	34	6	-	-	14	11	3
Mates	CW057	27	2	-	-	1	12	12
Mates	CW060	30	5	-	-	8	12	5
Mates	CW067	27	5	-	-	18	4	-
Mates	CW073	49	20	-	-	16	10	3
Mates	CW074	30	3	-	-	5	11	11
Mates	CW081	49	22	-	-	5	13	9
Mates	CW084	30	28	-	-	1	1	-
Mates	CW087	31	-	-	-	2	9	20
Mates	CW089	26	1	-	-	6	18	1
Mates	CW146	28	-	-	-	15	12	1
Mates	CW152	28	7	-	-	7	12	2
Mates	CW187	30	4	-	-	2	24	-
Mates	CW222	31	20	-	-	4	5	2
Mates	CW237	26	1	-	-	8	13	4
Mates	CW250	25	24	-	-	1	-	-
Mates	CW268	27	22	-	-	-	-	5
Mates	CW269	27	1	-	-	8	13	5
Mates	CW274	33	10	-	-	3	9	11
Mates	CW289	49	13	-	-	29	6	1
Mates	CW309	27	3	-	-	4	13	7
Mates	CW313	28	3	-	-	8	11	6
Nonmates	CW047	30	-	-	-	-	4	26
Nonmates	CW051	51	-	36	1	-	13	1
Nonmates	CW082	28	-	27	-	-	-	1
Nonmates	CW114	31	-	23	-	-	5	3
Nonmates	CW120	27	-	18	1	-	8	-
Nonmates	CW135	40	-	26	-	-	3	11
Nonmates	CW192	27	-	26	-	-	1	-
Nonmates	CW194	48	-	36	-	-	8	4
Nonmates	CW195	30	-	18	-	-	11	1
Nonmates	CW197	34	-	33	-	-	1	-
Nonmates	CW206	31	-	27	-	-	3	1
Nonmates	CW229	27	-	14	-	-	7	6
Nonmates	CW292	27	-	17	-	-	6	4
Nonmates	CW298	30	-	-	-	-	-	30
Nonmates	CW325	30	-	17	-	-	4	9
Nonmates	CW406	28	-	27	1	-	-	-
Nonmates	CW410	26	-	13	-	-	9	4
Nonmates	CW447	31	-	19	1	-	11	-
Nonmates	CW476	37	-	35	2	-	-	-
Nonmates	CW488	27	-	22	-	-	5	-

Table S2. Number of outcomes for each latent-exemplar image pair

Mating	*ImagePair*	*N Rows*	*N(TP)*	*N(TN)*	*N(FP)*	*N(FN)*	*N(Inc)*	*N(NV)*
Mates	CE011	47	47	-	-	-	-	-
Mates	CE012	30	30	-	-	-	-	-
Mates	CE013	26	26	-	-	-	-	-
Mates	CE014	25	25	-	-	-	-	-
Mates	CE015	25	25	-	-	-	-	-
Mates	CE016	30	30	-	-	-	-	-
Mates	CE017	29	29	-	-	-	-	-
Mates	CE018	31	31	-	-	-	-	-
Nonmates	CE001	35	-	34	1	-	-	-
Nonmates	CE002	48	-	48	-	-	-	-
Nonmates	CE003	31	-	31	-	-	-	-
Nonmates	CE004	26	-	26	-	-	-	-
Nonmates	CE005	25	-	25	-	-	-	-
Nonmates	CE006	26	-	26	-	-	-	-
Nonmates	CE007	27	-	27	-	-	-	-
Nonmates	CE008	26	-	25	1	-	-	-
Nonmates	CE009	29	-	28	1	-	-	-
Nonmates	CE010	34	-	33	1	-	-	-

Table S3. Number of outcomes for each exemplar-exemplar image pair. Dash indicates zero outcomes.

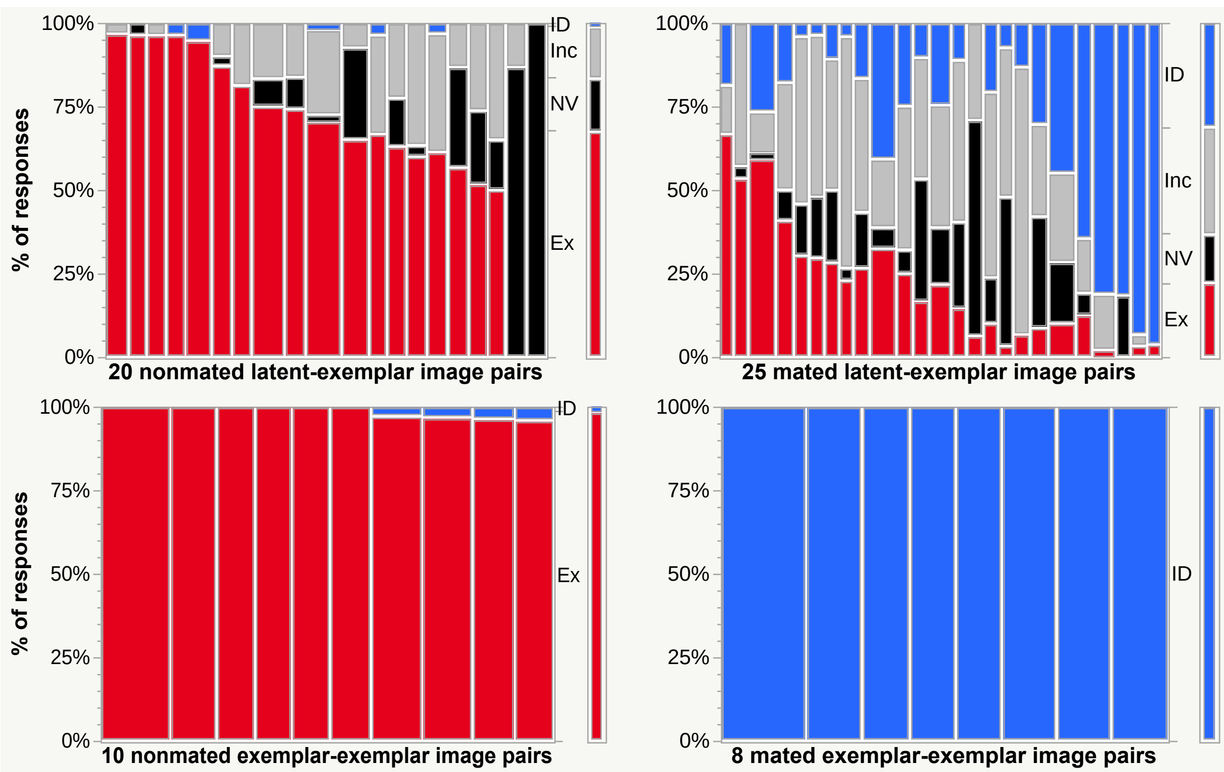


Fig S2. Distribution of conclusions for each image pair. Each column summarizes responses on one image pair, shown sorted by responses with width proportionate to the number of responses; multiple colors in a single column indicate the mix of responses by all examiners on that image pair. (n=804 mated and 640 nonmated latent-exemplar trials).

Fig S2 illustrates the distributions of conclusions for each image pair, and Tables S2 and S3 illustrate the number of outcomes for each latent-exemplar and exemplar-exemplar pair.

Appendix SI-3 Metric development, test instructions, and additional analyses

Appendix SI-3.1 Metric development

Appendix SI-3.1a Fixation segmentation and drift correction

The kilohertz eye-gaze stream was segmented into fixations and saccades. We used the Engbert–Mergenthaler saccade detector [3] with a velocity threshold of 8 pixels/second (0.16°/second) and a saccade duration minimum threshold of 9 ms to identify long saccades. As in Port, Trimberger [4], a saccade required that the eye remain at rest for at least 20 ms within a ± 0.25° X-Y positional window. Further details are found in Hicklin, Ulery [5].

Drift correction was estimated and corrected using an automated procedure that used a series of 9 fixation dots that began each trial. Examiners were asked to fixate each dot, and post-processing procedures estimated the horizontal and vertical transitions of the eye-gaze data to bring the fixation in correspondence with the fiduciary points. In addition, on trials where the Analysis phase included the placement of at least four marks, we took advantage of the fact that the act of precise placement of a mark on a minutia requires extensive fixation behavior. This affords an opportunity to make small manual corrections to the vertical and horizontal placement of the points so that they align with marks placed by examiners. This procedure was done rarely and conservatively, with the goal of reducing drift while not introducing systematic bias or distortion in the fixation points.

Once we segmented the gaze stream into fixations and saccades, we computed statistics such as the proportion of fixations on the latent vs the exemplar images, the number of fixations prior to a saccade to the other impression, as well as statistics such as the spatial dispersion of the fixations. These serve as proxies for cognitive capacities such as the capacity and duration of visual working memory (as estimated by the number of fixations prior to a saccade to the other impression), and the thoroughness of the search process (as estimated by the standard deviation).

Appendix SI-3.1b Fixation spread

Fixation spread is one measure of the degree to which examiners explore the available information. We computed the combined standard deviation in the horizontal and vertical dimension for the fixations on each trial:

$$std.dev. = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[(x_i - \bar{x})^2 + (y_i - \bar{y})^2]}$$

where x_i and y_i correspond to the x- and y-coordinates of the i^{th} fixation on the image, $\bar{x}$ and $\bar{y}$ are the average x- and y- coordinates from the trial, and n is the total number of fixations on the image.

This was done separately for the latent and exemplar images, and for both the Analysis and Comparison phases. The results in Fig 8 and Fig 9 in the main paper are for the latent impressions during the Comparison phase, but results are similar on the exemplars and during the Analysis phase.

Appendix SI-3.1c Subphase determination

The goal of subphase determination is to label fixations based on our inference of the underlying cognitive behavior, with a particular focus on detailed fixation that we assume are the basis of deliberative comparison behavior. By labeling a fixation as either scanning or detail, we intend to filter the fixations to focus on behavior that is central to the comparison task.

Kundel [6] decomposed eye-gaze behavior into scanning, recognition, and deciding subphases. Hicklin, Ulery [5] adapted this approach for latent print comparisons by developing an analysis rubric based on the time taken to fixate a target feature in a simple find-the-target task. In the three subphases described in Hicklin, Ulery [5], subphase A (scanning) is the period prior to finding the target, subphase C (deciding or detailed comparison) is the period after which the examiner has found the target and has stopped considering areas outside the target, and subphase B is any time not in A or C. Subphase C involves detail-like fixation behavior associated with back-and-forth comparison of specific areas and/or studying a location in the exemplar to determine if it corresponds to the memorized fingerprint region.

That approach took advantage of the fact that the target location on the exemplar was known. However, in the present work, we do not know where the target feature is on the exemplar. An open-ended comparison (as in this study) involves a series of comparisons of target groups [7], each of which may include a series of scanning and deciding subphases Hicklin, Ulery [5] included models using logistic regression to predict subphases. Here we built on that work by using machine-learning approaches trained on the find-the-target tasks to label the analogous subphases in these naturalistic comparison trials. This classification used AdaBoostM1 with a J48 classifier, and had an 84.71% accuracy on a cross-validation set with 10 folds. Training was conducted on fixation data from [5] (4959 subphase A fixations and 9694 subphase C fixations, omitting ambiguous or subphase B fixations, on 675 trials by 117 examiners).

Once trained, this model was then used to label the fixations in the current work, using the labels *scanning* and *detail* fixations to be more readily interpretable than subphases A and C.

Several of the features used in this classification process are described as "metrics of overt behavior" in Hicklin, Ulery [5] (copied verbatim here):

- *Speed3* measures the speed of eye movement within an image in image pixels per second. *Speed3* is measured over a series of up to ± three fixations, as the sum of inter-fixation distances, divided by the time from the start of the first fixation to the end of the last fixation in the series. Fixations near left-right transitions will have fewer than ± three fixations in the series.
- We use "image visit" to refer to a consecutive series of fixations in an image (left or right), from the start of the first to the end of the last fixation in that image, bounded by the transitions between the two images.
  - TimeInImage is the duration of time spent in each image visit.
  - DetailedBackAndForth is a count of image visits in which the examiner returns to the same small area in consecutive image visits in that image. The examiner is considered to have returned to the same small area if the distance between the centers of mass of consecutive image visits in the same image is within 88 pixels (approximately 4 ridges), and the max distance between any two fixations in each image visit is no more than 176 (2*88) pixels.

In addition to these, we used two additional eye-gaze metrics:

- *DistanceOfPriorSaccade* measures the distance of the saccade preceding the current fixation (i.e., the distance from the previous fixation).
- *FrequentBackAndForth* is a count of image visits in which the examiner makes a series of saccades between the two images, spending two seconds or less in each image.

Appendix SI-3.1d Spatial clustering

Within the eye-tracking literature, images are often segmented into regions of interest. Normally this assignment of regions of interest would be done by making assumptions about the relevant distinct features in an image; for example, with faces we would assume that the eyes, nose, mouth, and ears are all regions of interest.

However, fingerprints lack such well-defined regional features, which is to say that it is difficult to draw boundaries around informative features. This is not to say that all regions are equally informative, because the distribution of fixations across the images is quite 'clumpy'. However, the boundaries between regions are not well-defined and it is unclear how one might segment visual information in a fingerprint. As an alternative, we chose to use examiner-specific regions of interest. We clustered our fixations using a mean-shift clustering algorithm [8]. We submitted all of the fixations for the latent to this algorithm to determine both a set of cluster centers, as well as those fixations that belong to each cluster; the same process was applied separately to the fixations associated with the exemplar. Note that the clustering solution uses data for only that trial; we did not combine data across examiners for this clustering solution. In addition, our use of the term 'region of interest' refers to local regions of high fixation density, not the use in fingerprint comparisons where region of interest refers to the continuous area being compared.

This algorithm has the advantage that the number of clusters does not have to be predetermined. Instead, the algorithm uses a spatial 'bandwidth' that specifies how far apart we expect two regions of interest to lie. The algorithm uses this bandwidth to cluster based on density until it reaches a region of maximum density given the spatial bandwidth parameter. An analogy would be a mountain climber who always moves up the mountain, but will move down slightly if they can see a higher point within their field of view. The bandwidth is essentially how far the climber can see, and extent of the cluster is represented by the foothills and approaches of each peak. In our application, this algorithm places cluster centers at local regions of greatest fixation density that are at least one bandwidth apart by starting at a random fixation and iteratively moving toward the point of maximum density. All of the fixations that lead to this local density maximum are labeled as associated with this cluster. We selected a bandwidth of 66 pixels, corresponding to 3 average ridge widths, (Peak-to-peak ridge distance varies by subject and is affected by distortion, but the mean in criminal databases for males is 0.022" (0.56mm) [9], so at a resolution of 1000 pixels per inch (39.37 pixels per mm), 22 pixels is a useful heuristic for ridge width) which is the only parameter for this clustering algorithm. Fig 1 in the main paper illustrates an example of clustering. Larger colored circles with red borders are cluster centers, whereas smaller colored points are the fixations associated with that cluster. Note that in many cases the clusters may have only one or two fixations and are masked by the larger symbol.

Appendix SI-3.1e Temporal transition assignment and correspondence attempt estimation: The TECA model

A foundational element of a latent print comparison task is attempts at establishing correspondences between similar regions on the two impressions (examiners often describe two regions as being 'within tolerance'). We introduce a novel measure that is intended as a proxy for this behavior. The goal is to estimate the correspondence attempts made by examiners and relate this to outcomes. Our approach will eventually assign a set of correspondences between clusters of fixations on the latent and those on the exemplar.

To address this goal, we developed a novel approach we term the TECA model, for Temporal Estimates of Correspondence Attempts. This model is inspired by work in the concept literature [10] and has the goal of jointly assigning a set of correspondences while respecting a variety of constraints. In the TECA model, we rely solely on temporal information, in that the spatial relationship of one cluster to another is done prior to estimating the temporal correspondences. We explored other approaches that include spatial information, but ultimately settled on the TECA architecture for parsimony.

Correct identification decisions may be associated with more comparison attempts relative to erroneous exclusions, and the TECA model estimates these correspondence attempts. We refer to these as 'attempts', because presumably non-mated impressions would not actually have corresponding regions, yet this behavior is still likely to be an important part of the comparison process.

Inferring correspondence attempts from the eye-gaze record is complicated by the fact that not all information is captured by an eye tracker. We know where the eyes point, but not what information is acquired, nor can we directly measure the contents of visual working memory or information acquired through the visual periphery. Examiners report placing a small area containing several minutiae into working memory (referred to as a 'target group'), making a saccade to the exemplar, and searching for a visually similar region. The relative impermanence of visual working memory requires a saccade back to the latent, typically after an average of 5-10 fixations on the exemplar. All of these behaviors are readily observed with an eye tracker.

The clustering described in Appendix SI-3.1d is done separately for the latent and exemplar images, and therefore the cluster solutions represent potential areas of interest. To establish potential correspondence links, we computed the temporal transitions between the clusters in latent and exemplar images. This relies on the supposition that the constraints of visual working memory and the need to foveate the small details in minutiae requires first scanning for and memorizing a target group in the latent, followed by a rapid saccade to the exemplar to identify a possible correspondence. It is this temporal congruity that we wish to capture in a temporal transition.

The TECA model relies on temporal transitions from the latent image to the exemplar image. These transitions are sequences in the eye-gaze record that consist of a series of fixations on the latent followed by a series on the exemplar.

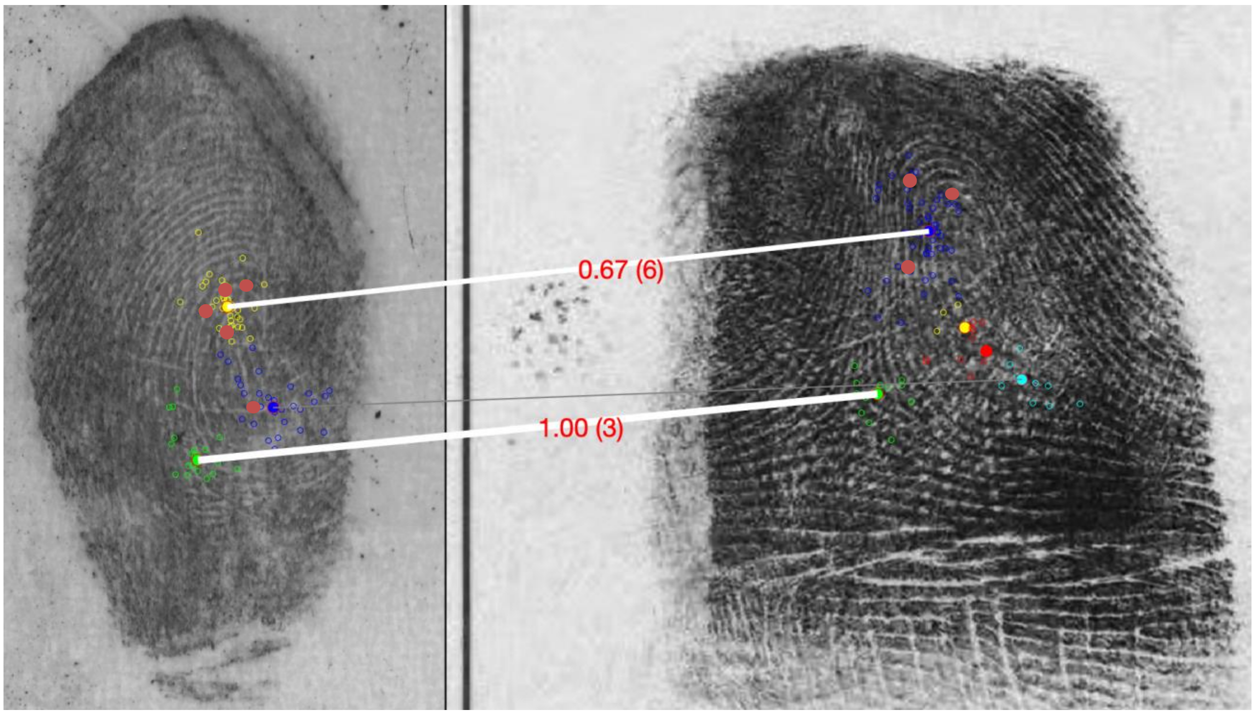


Fig S3. Illustration of a temporal transition sequence that forms the bases of the TECA model. In this hypothetical example, an examiner has made five fixations on the latent (four on the yellow cluster and one on the blue cluster, as shown by the dark red dots). The examiner then made a saccade to the exemplar, where she/he placed 3 fixations on the blue cluster. The temporal transition matrix would be updated one time for the yellow-blue pairing and one time for the blue-blue pairing. Subsequent latent-exemplar transitions will build up the entire temporal transition matrix, and there were 6 separate temporal transitions between yellow on the latent and blue on the exemplar (5 in addition the one described above). The white lines illustrate the correspondence attempts that were estimated by the TECA model, the red numbers are the relative strengths after normalization, and the numbers in parentheses are the number of temporal correspondences that contributed to each correspondence attempt. In this case, the green-green correspondence had fewer transitions (3) than the yellow-blue (6), yet had a higher relative strength after normalization because this subject was more definite when making green-green transitions (i.e., only went to green on the exemplar after fixating the green cluster on the latent).

We accumulate the identities of all of the clusters that were visited on the latent and the exemplar as part of that temporal transition. We create a *temporal transition matrix* that stores the number of times each cluster on the latent was associated with each cluster on the exemplar through a temporal transition, as shown in the example in Fig S3. Each time two clusters are associated via temporal transition, the cell of this matrix corresponding to the fixated latent and exemplar clusters is incremented by one. Note that multiple clusters could be associated during each temporal transition, because the examiner may fixate multiple clusters on the latent prior to a transition to the exemplar, and it is not clear that the last area they fixate on the latent is the feature they are searching for in the exemplar. However, our approach is designed to associate most latent clusters with a cluster on the exemplar.

Once the temporal transition matrix is populated, we use it to assign correspondence attempts between latent and exemplar clusters subject to the following constraints:

1) We first normalize the rows and columns of the temporal transition matrix by dividing each cell in a row by the row total, dividing each cell in a column by the column total, and averaging these normalized values. This normalization serves to emphasize latent-exemplar cluster pairs that were specifically looked at together. This gives fairer weight to correspondences that required few transitions and rewards 'monogamy' between two clusters. It also has the effect of establishing a weight value that ranges from 0 to 1 for all correspondences, which is important for the threshold operation described in 3) below.
2) We assign a unique correspondence for every latent cluster onto a cluster on the exemplar (unless there are more latent clusters than exemplar clusters, in which case some latent clusters will be unassigned). This assignment uses a greedy max algorithm, such that the largest cell in the normalized temporal transition matrix is assigned a correspondence, and then the rows and columns of the assigned cell are set to zero. This process is then repeated, always taking the max of the remaining entries in the temporal transition matrix to assign the next correspondence attempt.
3) Not all correspondence attempts are equally strong, and we eliminated weak correspondence attempts that fell below a threshold value of 0.3. This value was chosen based on introspection, but other values could be adopted without significantly affecting our results. The remaining high-quality correspondence attempts are

our estimates of where an examiner attempted to find 'detail in agreement' between the two impressions (to use a term favored by examiners).

To illustrate the TECA model in practice, take the following contrived temporal transition matrix (with letters arbitrarily assigned as cluster references). The numbers below represent the number of times the examiner placed a fixation in a latent cluster (A, B, or C), followed immediately by a fixation in a cluster on the exemplar (D, E, or F).

$$\begin{array}{cccc} & A & B & C \\ D & 5 & 25 & 7 \\ E & 20 & 30 & 10 \\ F & 8 & 35 & 25 \end{array}$$

Each temporal transition might associate several clusters at the same time; if an examiner looked at both clusters A and B followed by a saccade to cluster D, both the A/D cell and the B/D cell would be incremented in temporal transition matrix.

The TECA model uses information from the entire matrix to disambiguate the temporal transition matrix. For example, note that cluster B is looked at often, and although it is associated with D, E, and F, none dominates. However, A is associated with E, and C with F. We can use this information to help assign a correspondence for latent cluster B, and we do this in part using normalization by dividing each cell in a column by the column total, and each cell in a row by the row total. Accordingly, the above matrix would be normalized by column:

$$X = \begin{array}{cccc} & A & B & C \\ D & 0.152 & 0.278 & 0.167 \\ E & 0.606 & 0.333 & 0.238 \\ F & 0.242 & 0.389 & 0.595 \end{array}$$

and by row:

$$Y = \begin{array}{cccc} & A & B & C \\ D & 0.135 & 0.676 & 0.189 \\ E & 0.333 & 0.500 & 0.167 \\ F & 0.118 & 0.515 & 0.368 \end{array}$$

and the two would be averaged elementwise to produce a final matrix, representing a score between each cluster pairing:

$$\frac{X+Y}{2} = \begin{array}{cccc} & A & B & C \\ D & 0.143 & \mathbf{0.478} & 0.178 \\ E & \mathbf{0.470} & 0.417 & 0.202 \\ F & 0.180 & 0.452 & \mathbf{0.481} \end{array}$$

The greedy algorithm is used to assign the final correspondences. In this example, clusters C and F would be matched first (0.481), then B and D (0.478), and finally A and E (0.470). Since none of these values are less than 0.3, they would all be kept as correspondence attempts. Note one important detail about this assignment: Cluster B on the latent is assigned to Cluster D on the exemplar, despite the fact that it had only 25 correspondences with D in the raw temporal correspondence matrix, and 30 and 35 with the other two clusters. However, the constraints provided by the entire matrix help disambiguate individual correspondences.

Fig 1 in the main paper illustrates an example set of estimated correspondence attempts for a mated pair, and Fig S4 illustrates additional examples. We can assess the effectiveness of TECA in two ways: We can associate the number of correspondence attempts with the examiners' conclusions, and we can evaluate the correspondence attempts with respect to consensus minutia markup to identify the overall spatial accuracy of each correspondence attempt, as described next.

On most of our latent-exemplar mated pairs we have corresponding minutiae marked by multiple examiners, which were used to define consensus corresponding minutiae [11]. Based on these consensus correspondences, we used the thin plate spline (TPS) deformation model described in [12] to estimate corresponding locations between the images for all fixations. For example, using the TPS deformation model, for every fixation in a latent we estimated the corresponding location in the accompanying exemplar.

Using this TPS deformation model, we projected the region of maximum density of each latent cluster (essentially its centroid) onto the exemplar, which we treated as the "correct" corresponding location on the exemplar. We also estimate the correspondence attempt location on the exemplar using the TECA model. We compute the accuracy of each attempt by computing the distance between the TECA estimated location and the location from the TPS model on the exemplar. These distances are shown as red lines in Fig 1 in the main paper, and we might expect that TP outcomes might have smaller deviations than missed IDs.

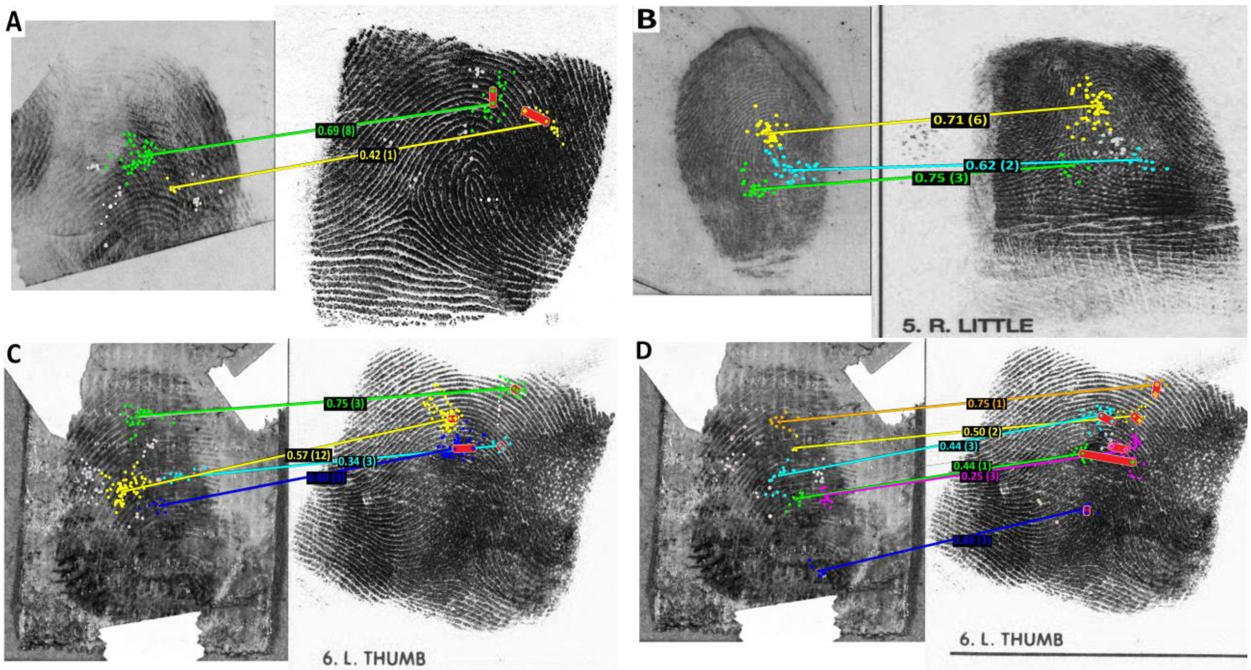


Fig S4. Additional examples of the TECA model applied to the detail fixations. Lines indicate correspondences, while red lines on exemplar images are the deviations between the estimated correspondence attempts and the ground-truth mated location of the latent cluster center. Gray lines are correspondences that were below strength threshold. Panel A: mated pair, ID conclusion, rated moderate difficulty. Panel B: nonmated pair, exclusion conclusion, rated difficult. Panel C: mated pair, ID conclusion, rated moderate difficulty. Panel D: mated pair, ID conclusion, rated very difficult.

Appendix SI-3.1f KS test statistic

Fig S5 illustrates the cumulative distribution of Comparison Time for TP outcomes, and the red vertical line is the KS test statistic. The associated p-value estimates the probability of a test statistic at least this extreme if the set of scores was sampled from the uniform distribution, which is the null hypothesis. This test is essentially asking whether a particular metric has an association with the ranks of a particular outcome, which would give a non-uniform distribution of outcomes at various levels of the metric ranks and therefore a deviation from the diagonal line. This test statistic can be computed for all of the ranks for a metric, or all ranks computed within an image pair. In either case, the collection of ranks associated with (in this case) TP outcomes is compared against the uniform distribution. Referencing the uniform distribution assumes that if there is no association, the ranks of the trials with TP outcomes would be uniformly distributed across the range from 0 to 1. To the degree to which the TP outcomes tend to be associated with higher ranks (either for ranks computed across all trials or ranks computed within an image pair), the cumulative distribution function (CDF) will deviate from the unform distribution CDF, and as shown in Fig S5, a TP CDF below the diagonal line is associated with TP outcomes tending to be clustered at the higher ranks. As this imbalance grows, the TP CDF will deviate more from the uniform distribution CDF, leading to larger KS test statistic values.

Because of the relatively large number of hypothesis tests we present, we will use a conservative alpha of 0.01 for our threshold for statistical significance, and describe results with p-values between 0.05 and 0.01 as weak evidence for an association.

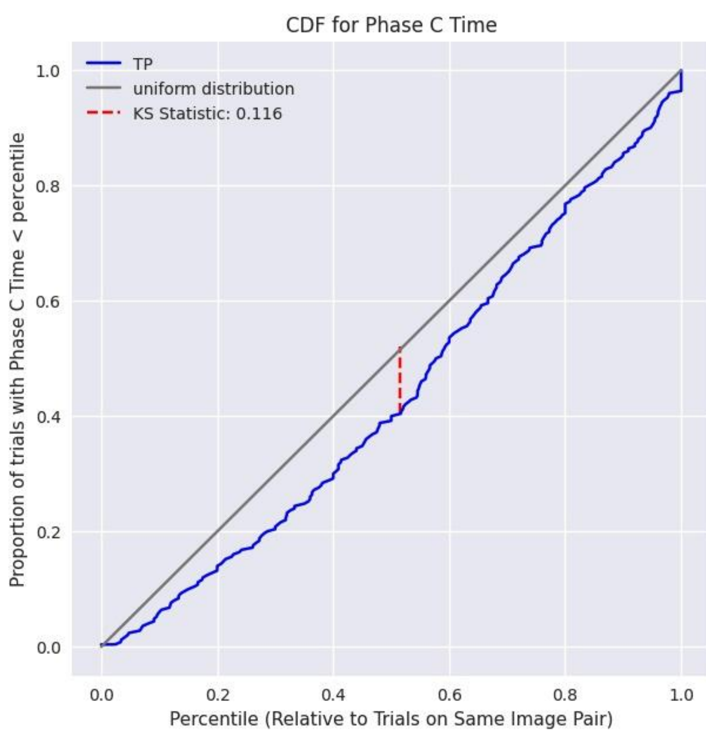


Fig S5. Illustration of the KS test statistic for the comparison of the CDF of the distribution of Ranked Comparison Time to the CDF of the uniform distribution (gray line). Ranked Comparison Time relative to trials on the same image pair is plotted for TP outcomes, and the dashed red line illustrates the maximum vertical deviation between the uniform distribution and the TP distribution. This is the KS test statistic, and it is calculated with the corresponding p-value using the Scipy package [13] to estimate the probability of obtaining a test statistic at least this extreme if the two sets of scores were sampled from the same underlying distribution. The generated graphs also used matplotlib [14], seaborn [15] and statsmodels [16].

Appendix SI-3.2 Summary of eye-tracking study test instructions

Ed. Note: this section is taken verbatim from the instructions provided to the eye-tracking study participants, but is limited to the content relevant to this paper (i.e., omitting directions regarding finding targets or following ridges).

In this study, you will be asked to perform a series of friction ridge impression examinations. Eye-tracking cameras will record the position of your head and eyes, in order to measure where on each image you are looking. Most of the test involves performing comparisons of two fingerprints. In addition, there are a few "directed tasks". It is important that you apply the same diligence that you use in casework when performing comparisons.

Analysis & Latent Value

For each examination, the software will first present a fingerprint at the left of the screen for analysis: generally, a latent, but occasionally an exemplar.

Once you complete the analysis stage, indicate the value of the print:

- *Of value for identification — The impression is of value and is appropriate for potential identification and/or exclusion if an appropriate exemplar is available.*
- *Of value for exclusion only — The impression is NOT of value for identification. The impression contains some friction ridge information that may be appropriate for exclusion if an appropriate exemplar is available.*
- *No Value — The impression does not contain sufficient friction ridge information to reach an identification or exclusion conclusion.*

Comparison/Evaluation Conclusion

At the end of comparison, you must make one of these conclusions: (say the conclusion aloud)

- *Identification — The two fingerprints originated from the same finger.*
- *Exclusion — The two fingerprints did not come from the same finger.*
- *Inconclusive — Neither identification nor exclusion is possible.*
- *Comparison not completed — You may choose this option if an examination is taking excessive time. You will be reminded of this option after 20 minutes, and asked to stop after 30 minutes.*

Borderline conclusion

Indicate whether your conclusion was a borderline decision, defined in this way:

- *If another examiner performed blind verification on this image pair and reached a different conclusion than you, how surprised would you be?*
    - Not borderline — You would be very surprised if another examiner disagreed: you would expect almost every qualified examiner to reach the same conclusion (say "NOT BORDERLINE")
    - Borderline — You would not be very surprised if another examiner disagreed: you would expect other examiners might disagree (say "BORDERLINE"). For Inconclusives, indicate "borderline ID" or "borderline exclusion."

[Why we ask: When assessing differences in eye behavior and differences in determinations, we want assistance in recognizing these borderline cases. For example, if you make an ID that is right on the edge of inconclusive, we want to be able to flag that as different from an ID you would expect every examiner to make.]

Using the software

The eye-tracking cameras that record the position of your eyes have limits on their accuracy. Because of this, we would like you to zoom in as much as you are comfortable when viewing the images.
To zoom*: move the mouse cursor over the image you want to zoom in or out. Press Z to zoom in, or X to zoom out. The zoom will be centered on the current mouse position. The two images need to be zoomed independently: you should always match the zoom level of the two prints. The two zoom levels are shown in the upper-left corner of the screen.*
To mark a minutia*: press the space bar and click on the location with the mouse.*
Discrepancies*: To indicate a minutia is a discrepancy, select the minutia and press D. The minutia will show in red.*
To link two minutiae*: select a minutia in the latent and a minutia in the exemplar and press L.*
To hide marked features*, press F.*
All other information you say aloud and the test administrator will record it.

Say "invert this image" to swap black for white in the image (tonal inversion).

Difficulty

For each comparison, say how difficult the comparison was. Routine comparisons should be indicated as "Moderate".

- *Very Easy/Obvious — The comparison determination was obvious.*
- *Easy — The comparison was easier than most latent comparisons.*
- *Moderate — The comparison was a typical latent comparison.*
- *Difficult — The comparison was more difficult than most latent comparisons.*
- *Very Difficult — The comparison was unusually difficult, involving high distortion and/or other red flags.*

Please Note

- *For the purposes of this study, "individualization" is treated as synonymous with "identification".*

- *Assume that the images provided are the only images available, and that physical evidence, lift cards, fingerprint cards, additional exemplars, and different images of these prints are not available.*
- *Every impression is a fingerprint, not a palmprint or lower joint.*
- *For an inconclusive determination, it is assumed that additional exemplars would have been requested and were not available.*
- *Make latent value determinations under the assumption that a good-quality exemplar with a large corresponding area may be available for comparison. The specific exemplar shown for comparison may or may not meet these criteria, but that should have no bearing on your value assessment.*
- *No images have already been claimed "Of Value." Some images were marked to indicate which impressions were to be scanned; such marks do not indicate that another examiner necessarily determined that the print is of value.*
- *All fingerprints are presented in an approximately upright orientation. The eye tracking software does not allow images to be rotated.*

Appendix SI-3.3 Analysis time

Fig S6 illustrates the boxplots for Analysis time broken down by different conclusions. This metric does not show a large association with different outcomes, although differences between images may mask smaller effects.

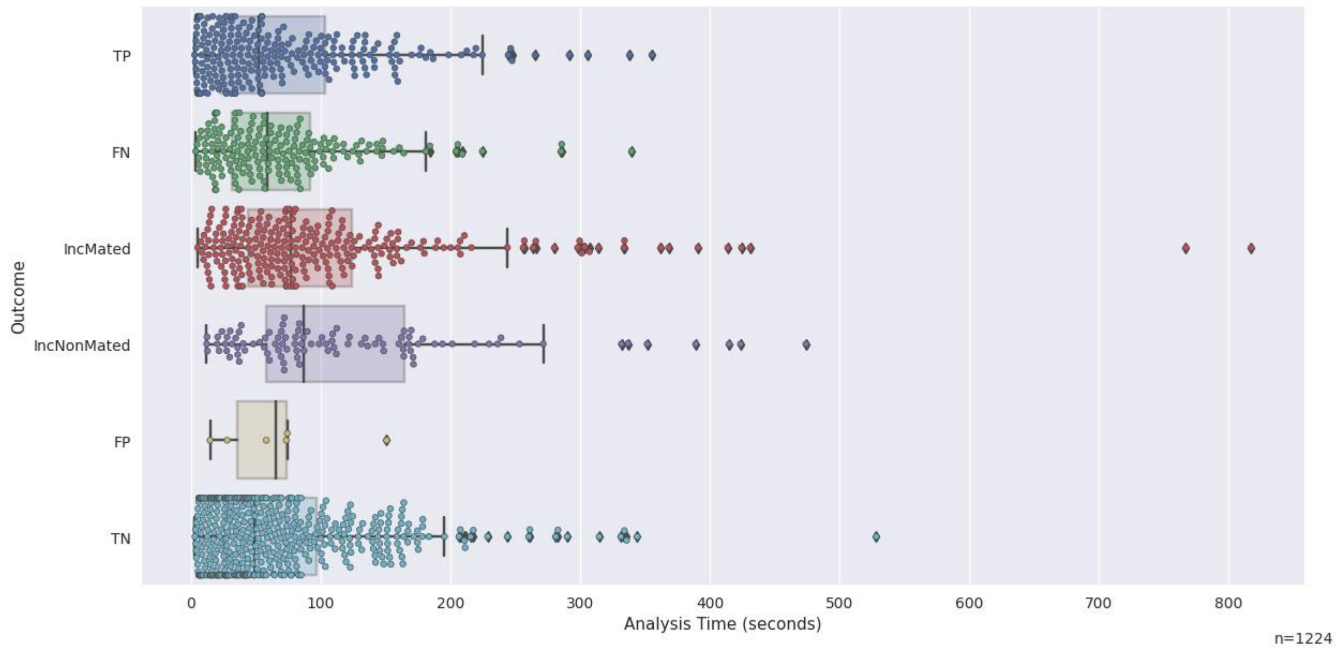


Fig S6. Distribution of Analysis Time grouped by outcome.

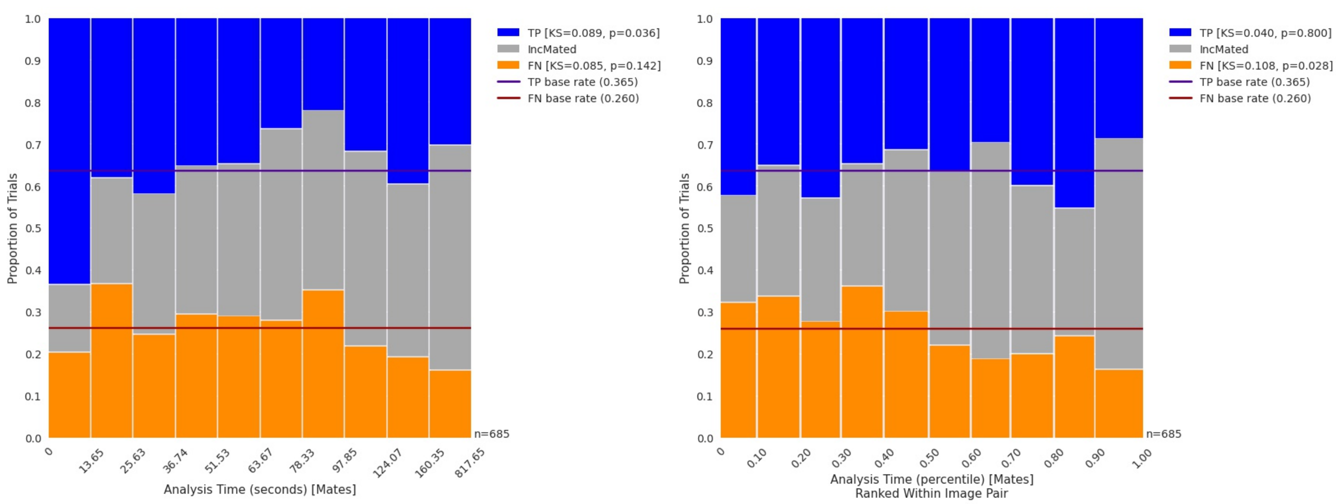


Fig S7. Distribution of different outcomes by Analysis Time, ranked by each image pair.

Fig S7. illustrates the Analysis Times ranked within each image pair. TP and TN outcomes do not show statistical support for a relation between outcome and Analysis time, with the exception of trials that had very short analysis times (left panel). These are likely due to easy images, because over half of the trials in the bottom decile in the left panel of Fig S7. are from just two images. Evidence for this association disappears once the trials are ranked by image pair (right panel of Fig S7.). FNs shows weak evidence for an association with Analysis Time.

Appendix SI-3.4 Number of fixations prior to switching images

The dynamics of eye gaze are fairly stereotyped for many tasks, not just for fingerprint comparisons: The eye tends to move (saccade) to a new position at a rate of approximately 3 times a second. However, the nature of a task and the ability of an examiner both contribute to the locations of those movements. One feature of expertise may be the ability to hold more visual features in working memory. We developed a proxy for the capacity and duration of visual working memory by counting the number of fixations on the exemplar prior to a saccade to the latent. We did not distinguish between latent and exemplar images when constructing this statistic to acknowledge the possibility that examiners may 'reverse compare' by looking for a target group on the exemplar and searching for it in the latent.

Fig S8. shows the boxplots for the number of fixations on the one image prior to a switch to the other image. TP appears to have slightly more fixations prior to a switch than FN or TN, although some images may have more regions to fixate and therefore image effects may drive this difference. These effects are modest at best, however.

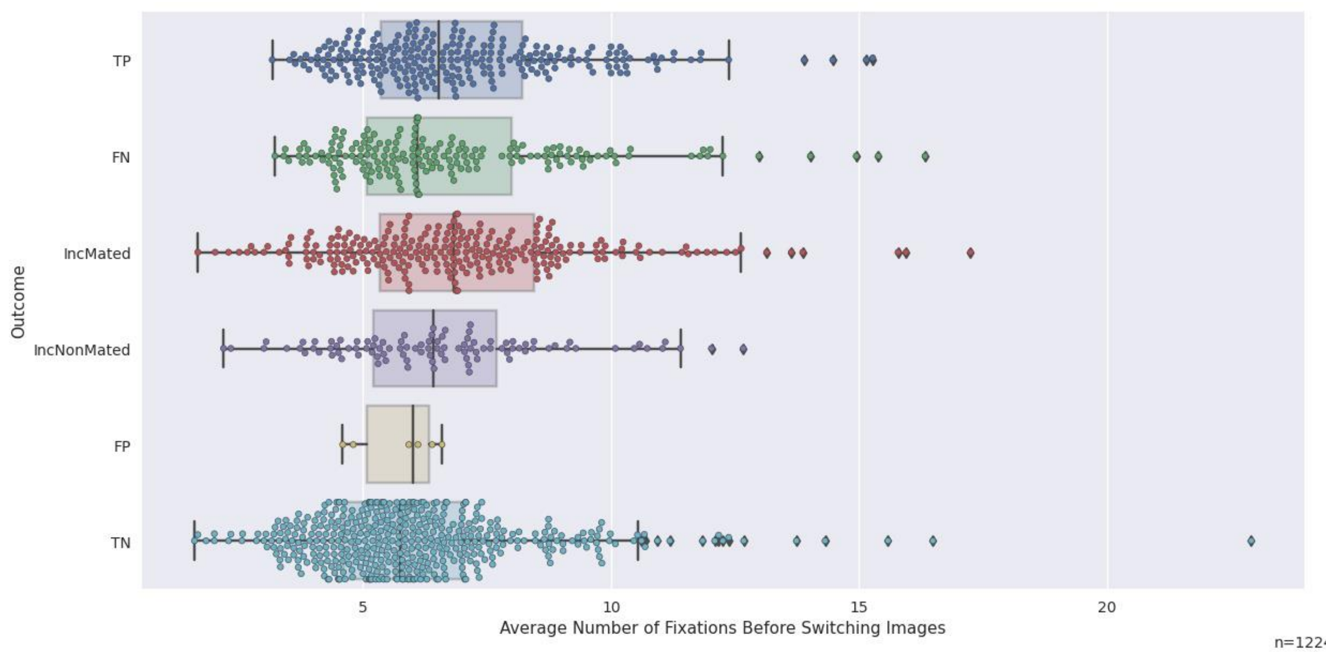


Fig S8. Average number of fixations prior to a saccade to the other image.

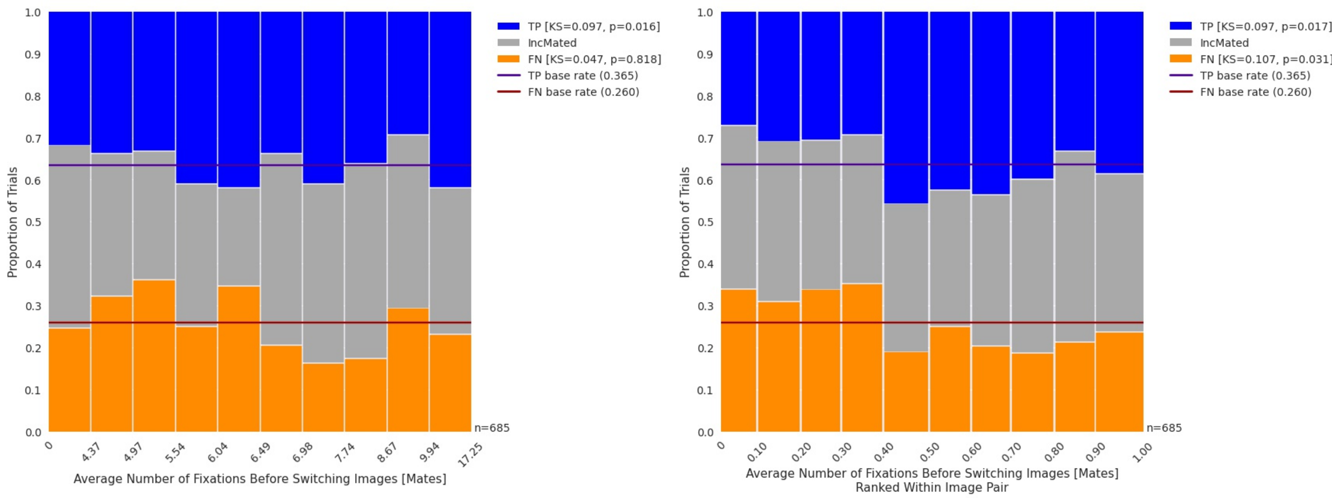


Fig S9. Distribution of different outcomes by number of fixations prior to a saccade to the other impression, ranked by each image pair.

The left panel of Fig S9 illustrates weak evidence for TP outcomes to be associated with more fixations prior to a saccade (KS = 0.097; p=0.016). The right panel of Fig S9 presents the histograms of this metric ranked within each image pair to reduce image effects. There is weak evidence for an association between TP and the number of fixations prior to a saccade (KS = 0.097; p=0.017), which does not reach statistical significance by our conservative alpha. We see a similar result between this measure and FN outcomes, with a modest tendency for more fixations to be associated with fewer FN outcomes.

The weak evidence for associations we observe may be a function of the familiarity that examiners have with the stimuli and, if we had included less experienced examiners or novices, we may have observed stronger evidence. However, this metric does not appear to be a major contributor to outcome.

Appendix SI-3.5 Image clarity

Examiners tend to describe a latent in terms of both the clarity of the print and the quantity of features present, and suggest that both play a role in the amount of information available to conduct a comparison. Examiners who are willing to look in lower-clarity regions may have different outcomes than examiners who rely primarily on higher-clarity regions. We used previously marked regions of low clarity (denoted by red and yellow in the clarity markup process) on many of our images to compute the proportion of fixations that fall in the lowest clarity regions [2].

Fig S10 illustrates that that there is perhaps weak evidence for higher numbers of fixations in low clarity regions for FN outcomes, but this could be due to image effects: Images that are of poorer clarity will lead to more fixations in poor clarity regions and also be more difficult. However, the data in Fig S10 is not consistent with the idea that better examiners are more willing to work 'in the weeds', or if so, that effect is overwhelmed by the image effects mentioned above.

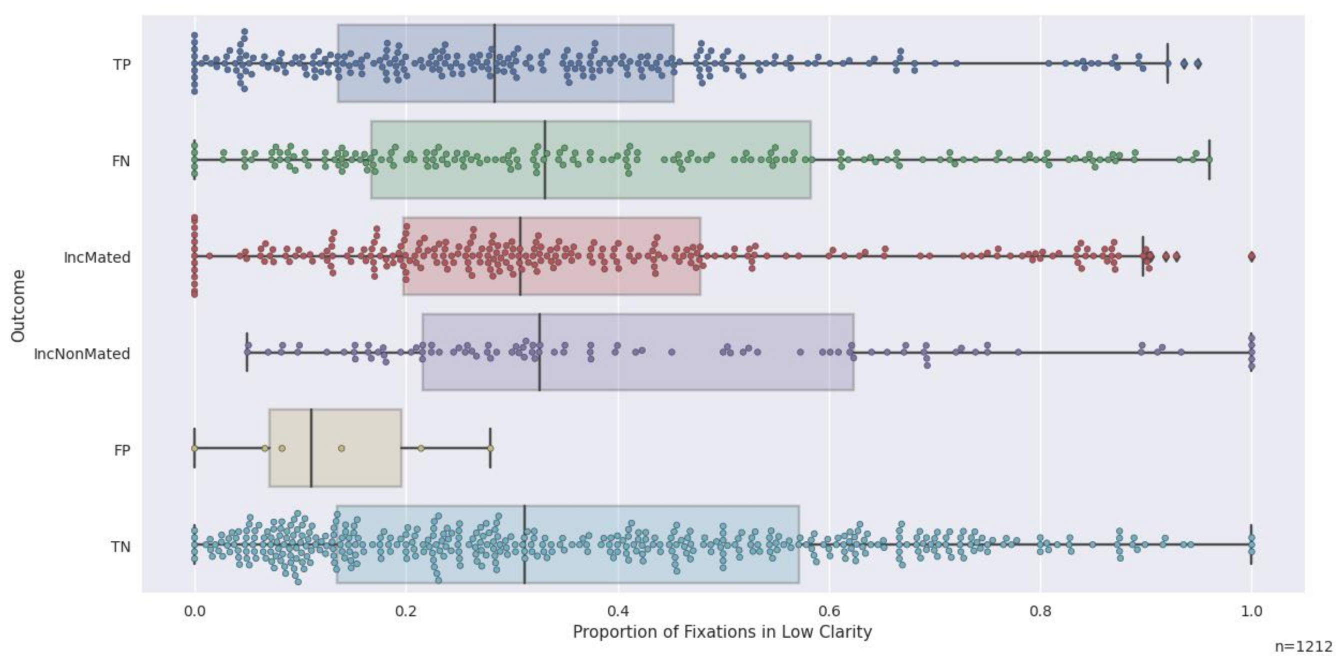


Fig S10. Boxplots of the proportion of fixations that fall in the low clarity regions of the latent impression.

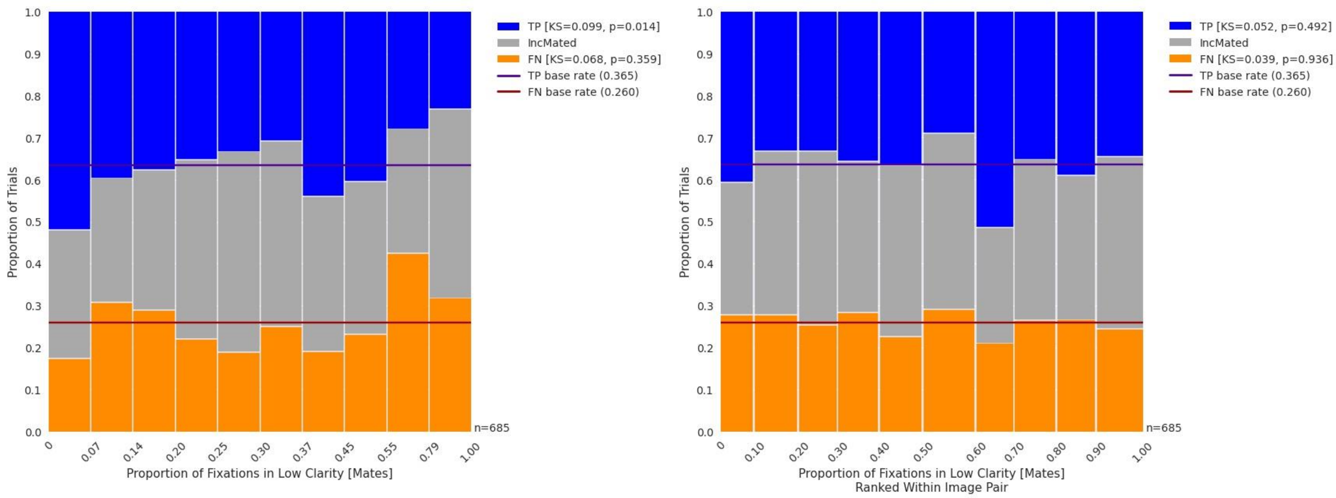
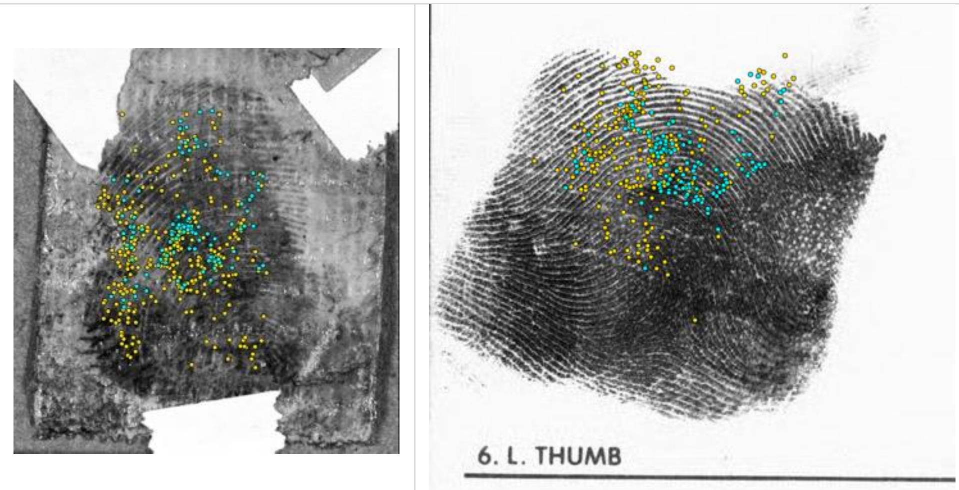


Fig S11. Relation between outcomes and the proportion fixations in low clarity regions. Left panel: data for the proportion of fixations in low clarity regions. Right panel: data the proportion of fixations in low clarity regions ranked within image pair.

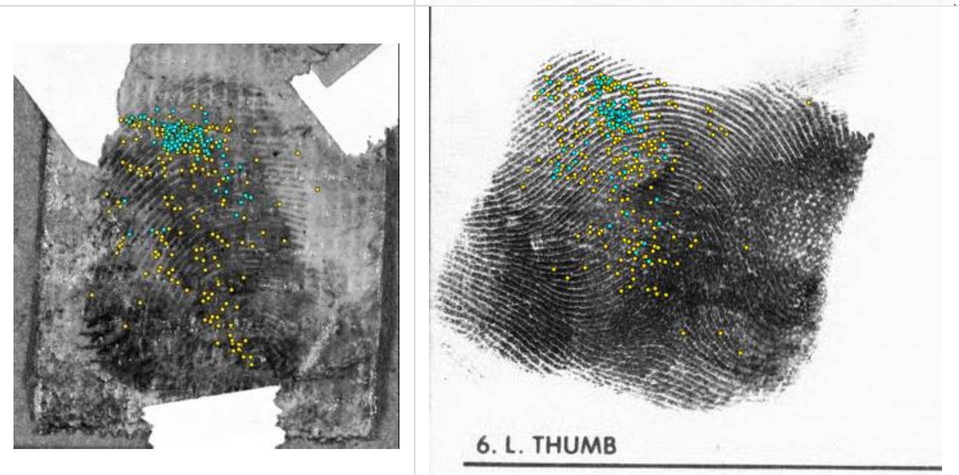
The left panel of Fig S11 shows that there is weak evidence for an association between TP outcomes and the proportion of fixations in low clarity regions. In this case, TP outcomes tended to be associated with *fewer* fixations in low clarity regions (KS = 0.099; p=0.014). However, this does not mean that one should never look in low clarity areas; it could be that images that only contain low clarity areas are more difficult. Once this metric is ranked within image pairs as demonstrated by the right panel of Fig S11, there is no evidence for an association between the proportion of fixations in low clarity regions and either TP (KS = 0.052; p=0.492) or FN (KS = 0.039; p=0.936) outcomes.

These results provide weak support at best for an association between the proportion of fixations in low clarity regions and TP outcomes, which are likely driven by image effects, not differences among examiners.

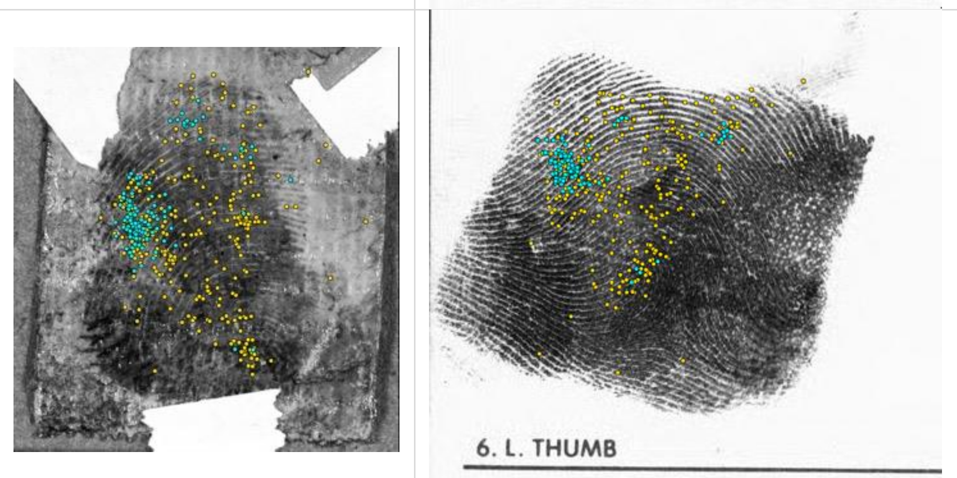
Appendix SI-4 Examples of fixation data



Top row:
Correct Identification
Moderate difficulty
4.9 minute comparison time

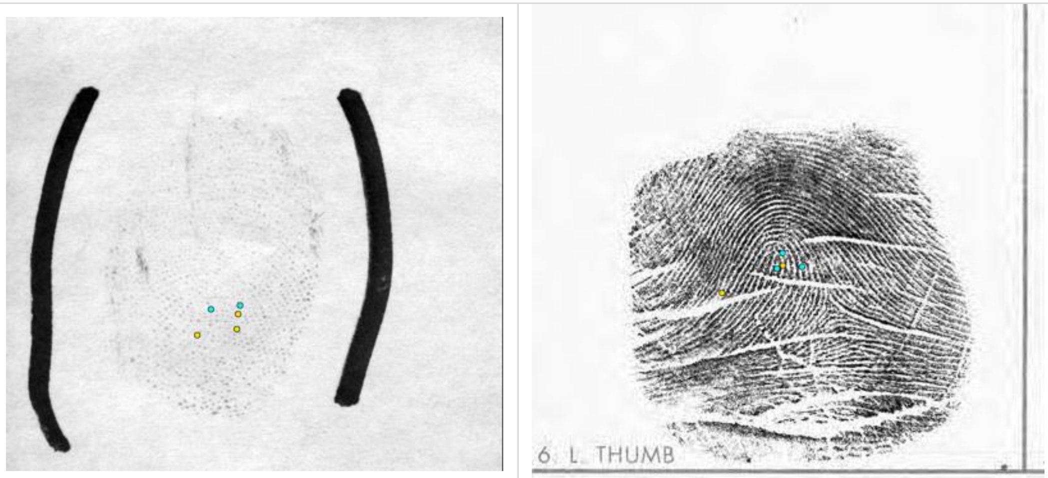


Middle row:
Erroneous Exclusion
Difficult
4.8 minute comparison time

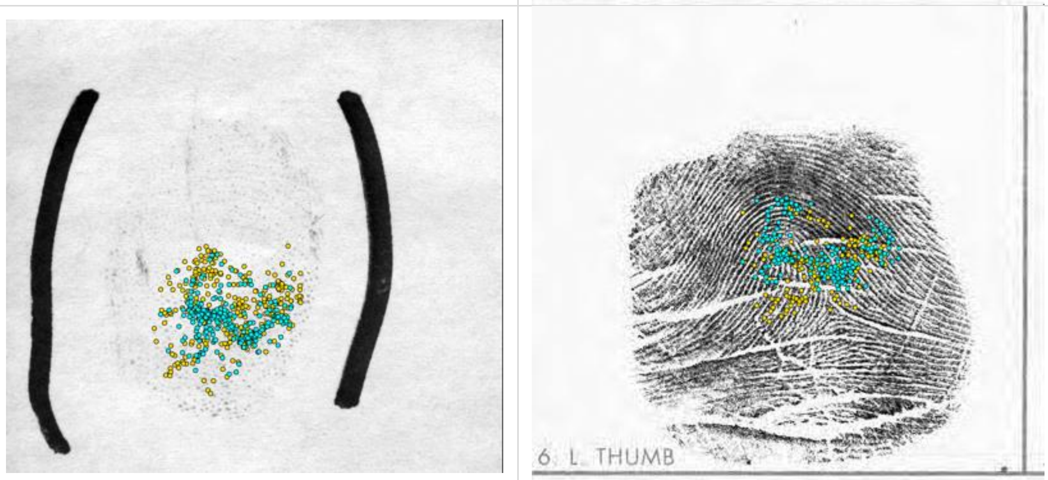


Bottom row:
Inconclusive
Difficult
5.2 minute comparison time

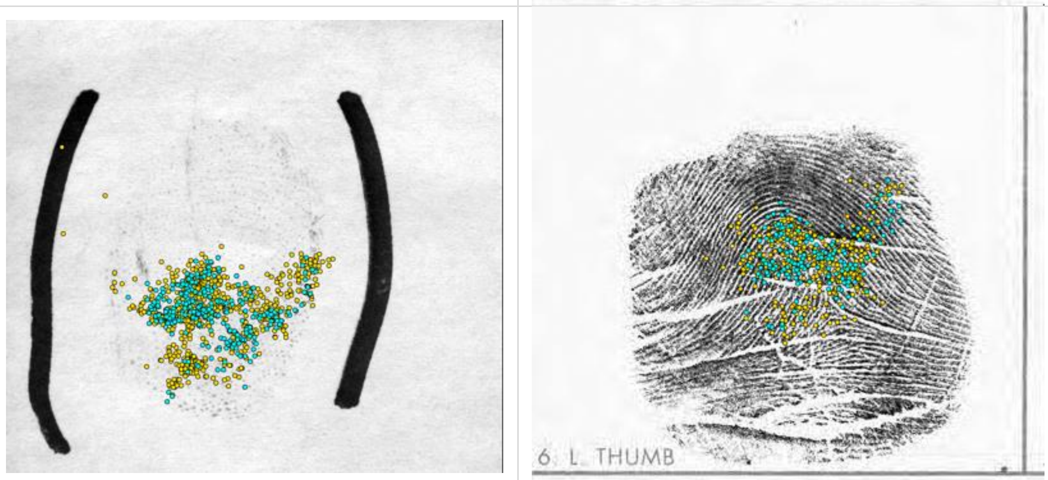
Fig S12. Examples of fixations for three trials on a mated image pair (CW073). Overall conclusion counts: No Value 3, exclusion 16, inconclusive 10, identification 20, Total 49.



Top row: Inconclusive
Very Difficult
5 second comparison time
(after 1.8 minute analysis time and determination of Value for exclusion only)



Middle row: ID
Very Difficult
7.2 minute comparison time



Bottom row: Inconclusive
Very Difficult
9.9 minute comparison time

Fig S13. Examples of fixations for three trials on a mated image pair (CW309). Overall conclusion counts: No Value 7, exclusion 4, inconclusive 13, identification 3, Total 27.

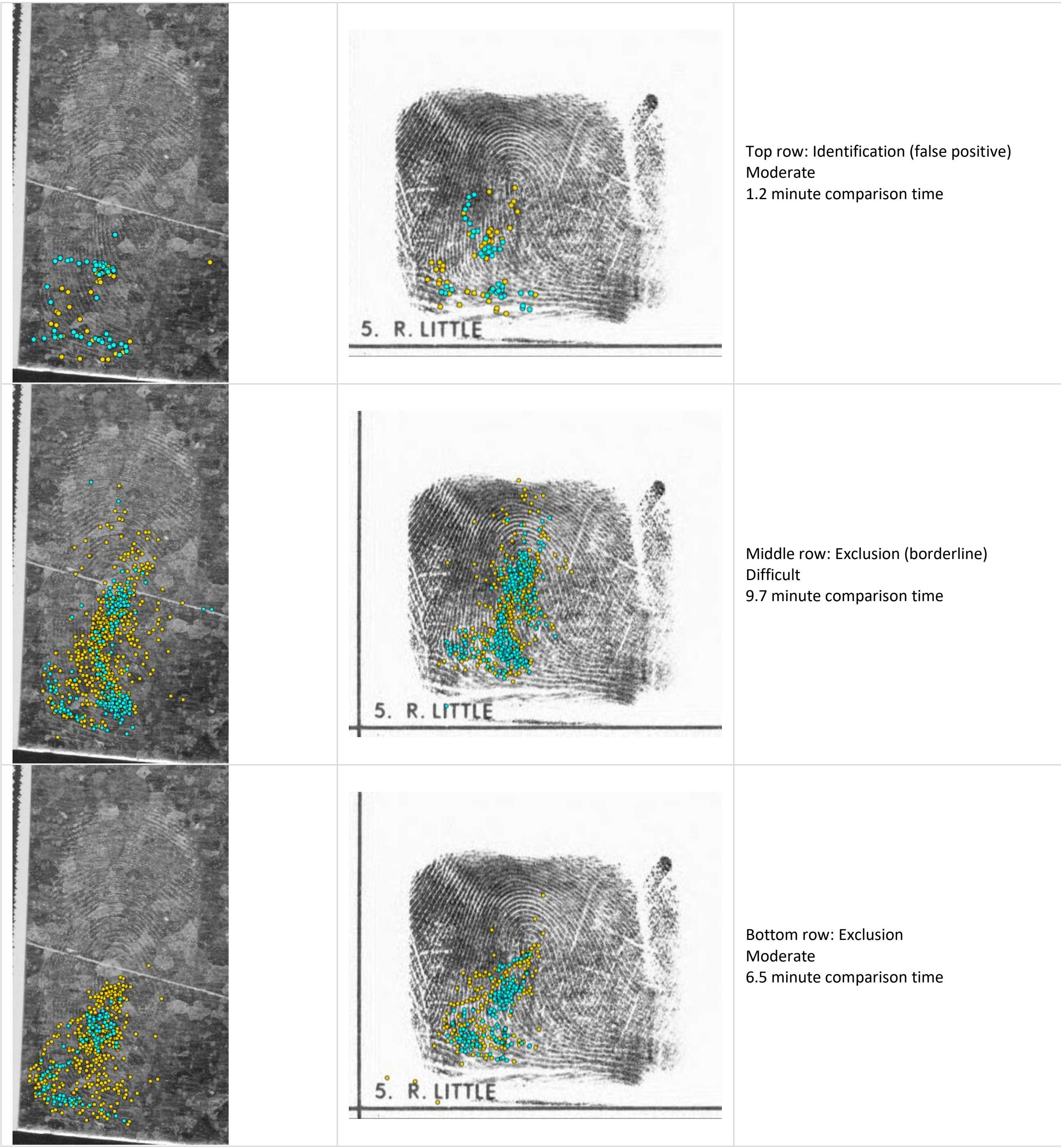


Fig S14. Examples of fixations for three trials on a nonmated image pair (CW447). Overall conclusion counts: Exclusion 19, identification 1, inconclusive 11, Total 31. The false positive trial is by an examiner other than the examiner discussed in Section 3.5 — none of the other false positive trials have releasable exemplar images (fingerprints are protected as Personally Identifiable Information and public release requires permission from the subject, which could not be obtained for those exemplars). Aqua points are detail/deciding fixations, and yellow points are scanning and miscellaneous fixations.

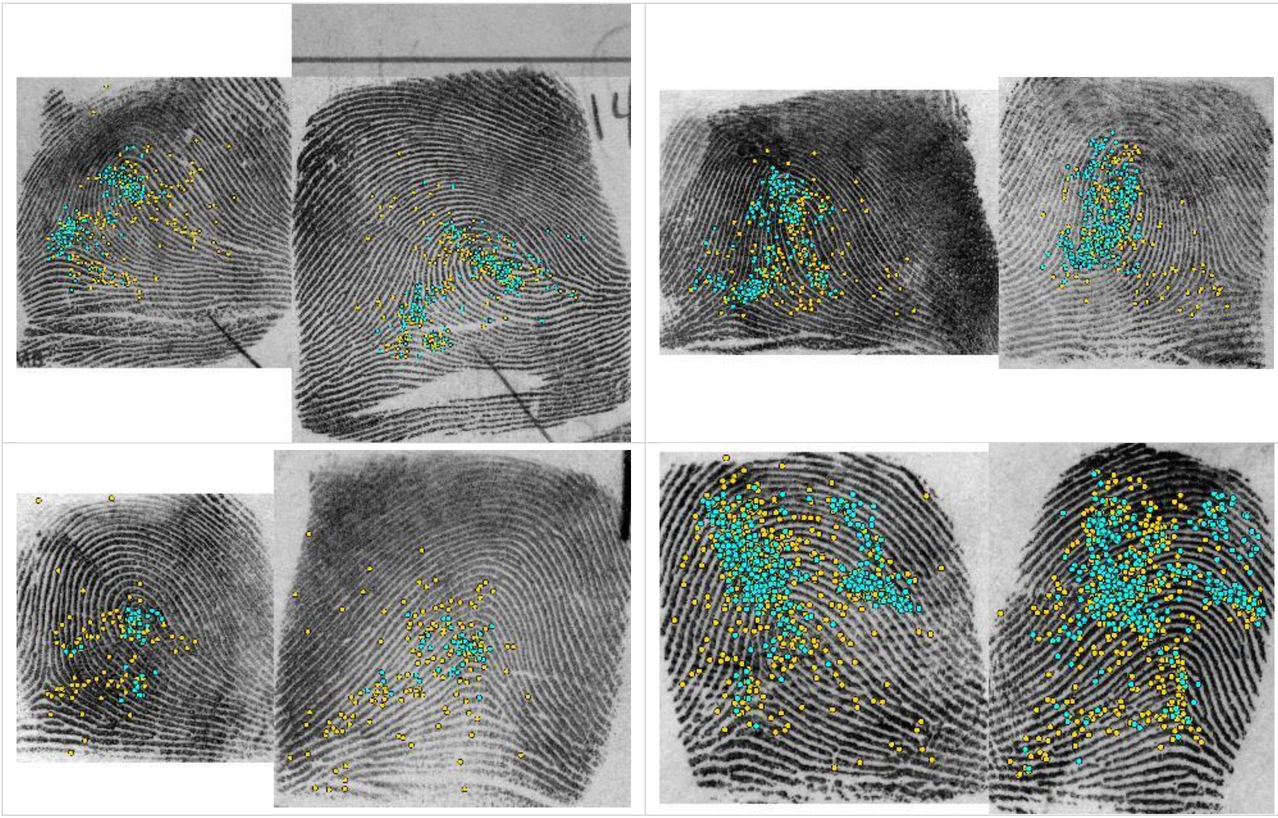


Fig S15. Raw fixations for four false-positive exemplar-exemplar trials by the participant discussed in Section 3.5; that examiner assessed the difficulty as moderate. No other participants made errors on exemplar-exemplar image pairs: all other conclusions on these image pairs were true negatives (exclusions) on which the difficulty was assessed as easy or very easy. Aqua points are detail/deciding fixations, and yellow points are scanning and miscellaneous fixations.

References

1. Ulery BT, Hicklin RA, Buscaglia J, Roberts MA. Accuracy and reliability of forensic latent fingerprint decisions. P Natl Acad Sci USA. 2011;108(19):7733-8. doi: Doi 10.1073/Pnas.1018707108. PubMed PMID: WOS:000290439500024.

2. Ulery BT, Hicklin RA, Roberts MA, Buscaglia J. Measuring what latent fingerprint examiners consider sufficient information for individualization determinations. Plos One. 2014;9(11):e110179. Epub 2014/11/06. doi: 10.1371/journal.pone.0110179. PubMed PMID: 25372036; PubMed Central PMCID: PMCPMC4221158.

3. Engbert R, Mergenthaler K. Microsaccades are triggered by low retinal image slip. P Natl Acad Sci USA. 2006;103(18):7192-7. doi: 10.1073/pnas.0509557103. PubMed PMID: WOS:000237399900073.

4. Port NL, Trimberger J, Hitzeman S, Redick B, Beckerman S. Micro and regular saccades across the lifespan during a visual search of "Where's Waldo" puzzles. Vision Res. 2016;118:144-57. doi: 10.1016/j.visres.2015.05.013. PubMed PMID: WOS:000368968600014.

5. Hicklin RA, Ulery BT, Busey TA, Roberts MA, Buscaglia J. Gaze behavior and cognitive states during fingerprint target group localization. Cogn Res. 2019;4. doi: ARTN 12 10.1186/s41235-019-0160-9. PubMed PMID: WOS:000463671000001.

6. Kundel HL, Nodine CF, Carmody D. Visual Scanning, Pattern-Recognition and Decision-Making in Pulmonary Nodule Detection. Invest Radiol. 1978;13(3):175-81. doi: Doi 10.1097/00004424-197805000-00001. PubMed PMID: WOS:A1978FE29600001.

7. Holder EH, Robinson LO, Laub JH. The fingerprint sourcebook: US Department. of Justice, Office of Justice Programs, National Institute of …; 2011.

8. Cheng YZ. Mean Shift, Mode Seeking, and Clustering. Ieee T Pattern Anal. 1995;17(8):790-9. doi: Doi 10.1109/34.400568. PubMed PMID: WOS:A1995RL03500005.

9. Ashbaugh DR. Quantitative-qualitative friction ridge analysis : an introduction to basic and advanced ridgeology. Boca Raton, Fla.: CRC Press; 1999. xiv, 234 p. p.

10. Goldstone RL, Rogosky BJ. Using relations within conceptual systems to translate across conceptual systems. Cognition. 2002;84(3):295-320. doi: Pii S0010-0277(02)00053-7 Doi 10.1016/S0010-0277(02)00053-7. PubMed PMID: WOS:000176306400003.

11. Ulery BT, Hicklin RA, Roberts MA, Buscaglia J. Interexaminer variation of minutia markup on latent fingerprints. Forensic Sci Int. 2016;264:89-99. doi: 10.1016/j.forsciint.2016.03.014. PubMed PMID: WOS:000378623200013.

12. Kalka ND, Hicklin RA. On relative distortion in fingerprint comparison. Forensic Sci Int. 2014;244:78-84. Epub 2014/09/13. doi: 10.1016/j.forsciint.2014.08.007. PubMed PMID: 25216456.

13. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature methods. 2020;17(3):261-72.

14. Hunter JD. Matplotlib: A 2D graphics environment. Computing in science & engineering. 2007;9(3):90-5.

15. Waskom M, Botvinnik O, O'Kane D, Hobson P, Lukauskas S, Gemperline DC, et al. mwaskom/seaborn: v0. 8.1 (September 2017). Zenodo, doi. 2017;10.

16. Seabold S, Perktold J, editors. Statsmodels: Econometric and statistical modeling with python. Proceedings of the 9th Python in Science Conference; 2010: Austin, TX.